MANAV RACHNA INTERNATIONAL INSTITUTE OF RESEARCH AND STUDIES

Short Synopsis

For

Ph. D. Programme 2012-13

Title: Design and Development of Efficient Technique for Classification in Data Mining

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

FACULTY OF ENGINEERING AND TECHNOLOGY

Submitted by:

Name: Preeti Nair

Registration No.: 12/Ph.D./007

Supervisor:

Name: Dr. Indu Kashyap

Designation: Professor

Abstract

Data mining refers to an automatic process of recognizing exceptionally useful and interesting patterns from huge databases. The major techniques of data mining are association rules, clustering and classification. Researchers have made immense advancements in data mining techniques in the past years where classification is possibly one of the most essential tasks and a widely studied research topic in the field of data mining. Classification predicts predefined class labels for unknown instances based on the behaviours of the past data.

This research work undertaken intends to develop an original and innovative classification technique, which shows enhanced performance for all types of datasets such as numeric, categorical, and mixed compared to the existing classification techniques in a multiplicity of data analysis environments.

k-Nearest Neighbor (k-NN) classifier has been a widely known classification technique. It is very simple and easy to implement. The basic idea behind k-NN is to find a relationship of a query instance to nearest neighbouring samples, which are measured using similarity matrix such as Euclidean distance or Manhattan distance etc. In this research, the main focus is to enhance the efficacy of k-NN classification technique to reduce gaps in its functional capabilities.

Some of the gaps in k-NN approach is that the value for k has to be given manually. Whenever a test instance is inputted for finding class label, k-NN calculates distances each time from each data point to the test instance in the feature space and this is a very time consuming task. The technique that will be proposed in this research would automatically find the optimal k value for a given dataset and it would able to handle all kinds of data such as imbalanced data and mixed data more efficiently. The proposed technique will also reduce the training phase cost and will be much improved in performance in terms of its accuracy.

Keywords: Classification, k-NN classification, optimal k value, imbalanced data, Outliers, distance measure, computational complexity, accuracy.

Contents

S. No.	Description	Pg. no.
1	Introduction	1-2
2	Literature review	2-3
3	Description of broad area	4
	3.1 Introduction to data mining	4
	3.2 Data mining techniques	4
	3.2.1 Frequent pattern mining and association	4-5
	3.2.2 Clustering techniques	5-6
	3.2.3 Classification techniques : k-NN Classifier	6-7
4	Open issues in k-NN Classifier	7-8
	4.1 Problem Identification	8-9
	4.2 Gaps in Literature	9-10
5	Objective of the study	10
6	Methodologies to be adopted	10-11
7	Expected outcome of the research	11
8	References	12-13

1. Introduction

One of the major challenges of information technology is the rapid expanse in the production of data. Almost all organisations such as banking, health sector, educational sector, manufacturing engineering, research fields, space study institutions etc are generating huge amount of data. This information is useful for taking right decision for the organisational growth only if the decisions are made correctly at the correct time. One of the ways for addressing this challenge is data mining: i.e., searching valuable insights from huge amount of data. So, Data mining is the process of discovering valuable information from vast amount of data. There are many predefined techniques available in this emerging field of data mining; some of the most widely used techniques are association rules, frequent pattern mining, clustering and classification.

Classification is a supervised learning method in which a classifier predicts the class label of an unseen instance. There are many domains in which classification is applied predicting credit risks, spam filtering etc. To classify a new data (unseen instance) for instance in a bank, a loan officer will analyse the data to learn which customer is "risky" or which is "safe" for the bank. A medical researcher would analyse a cancer data to predict which treatment would be beneficial for the patient (Han et al. 2012) (Kumar and Upadhyay 2015).

k-NN is a simple yet powerful classification algorithm. It is considered as a standard for many difficult algorithms such as neural network and support vector machine. Its various applications are cost-effective forecasting, data compression and medical data analysis etc. however, there are some limitations in k-NN. One of the limitations of k-NN is that its performance accuracy relies on the k value. k neighbourhood instances are selected around the query instance for classification. If the value of k is too small then the estimate tends towards noisy data, indistinct data and it results into a mislabelled classification. If k is too large, estimate tends to over smoothing outliers. The majority value of class present in the k neighbourhood will be the class of the query instance. So selecting the right k value for a given dataset is a challenge. Another limitation of k-NN is that it has a very expensive training phase. All the distances from the unseen query instance have to been taken first in order to find the closest distant neighbourhoods. To overcome these limitations this research work proposes a novel algorithm, which can find an optimal k value with more accurate classification results

with less expensive training phase (Larose, 2005). In order to increase the accuracy of the classifier, we have attempted to develop a hybrid classification technique based on k-NN (Nair et al. 2018).

2. Literature review

Aouidate, A et al. (2012) have proposed a method based on hybrid k-NN and ant colony optimisation. In this method, the instances are reduced in such a way that the accuracy of reduced training dataset is very close to the full set of training set. This method is called condensing. The condensed k-NN and ant colony optimisation together form a new approach which outperformed many other k-NN based approaches. The idea behind this method is valid however, upon closer study it is clear that the classification performance of this method can be improved further.

Zhang, S et al. (2012) have developed a new classifier called mutual k nearest neighbor classifier (MkNNC), which can remove the noise data. It can remove the anomalies and predict the class labels of unknown query instance. Rather than relying on the k-NN, it uses the mutual NN for class prediction; however, the decision for finding the class label of the query instance depends on a best "k" value. Therefore, finding the best value for k is very hard, if smaller "k" value is considered then noise sensitivity problem appears, if larger "k" is considered smoother decision boundaries and lesser noise problems emerge. This is the main issue countenanced with this classifier.

Gou, J et al. (2012) with aim of improving the classification performance have proposed a new method to find appropriate k value in this paper. A method called dual distance-weighted k nearest neighbor (DDWKNN) was introduced, which is based on weighted k nearest neighbor (WKNN). The experiment results showed larger values for k gave good performance. Nevertheless, the value of "k" for each dataset was given manually by users and not automatically generated. Another element which warrants more investigation is the accuracy aspect which we believe can be further improved.

Pedrajas, N.G. et al. (2015) have proposed a novel method for finding k value which is local to a specific region of the dataset. The performance of this method was much better than the conventional k-NN classifier. This method has shown good results for both standard datasets as

well as imbalanced datasets. In terms of accuracy, there is still a scope for improvement. Furthermore, this method can be extended by coupling the local k value with instance selection.

Yu, Z et al. (2015) have proposed a paper on special dataset problems such as imbalanced data, spare data, and have addressed using hybrid k-NN (HBKNN) to tackle the noisy data problem in the high dimensional dataset, an ensemble method for random subspace with HBKNN (RS-HBKNN) is proposed, which outperforms most classification approaches. However, in this method, an optimal value for "k" is not known and user has to input it manually.

Pan, H et al (2016) have proposed a new method multi local means based k harmonic nearest neighbor (MLMkHNN) based on the standard k-NN classifier which defeats the issues of uniform k values. This method calculates the harmonic mean distance to the unknown query. It showed that the proposed method has much less classification error as compared to other k-NN based classifiers. But still there is a scope for investigation to discover which appropriate distant measure can be used. It is improper when the distance measure of each nearest neighbor (NN) differs very much against the distance measure of the test data.

Zhang, X et al. (2016) has proposed new algorithm called k nearest neighbor which straightaway triggers the rare class problem by forming the dynamic query neighbourhoods and adjusting the positive posterior probability. This method has improved the k-NN classification on rare class problems. However further, this method can be investigated, like for example its efficacy in extending it to handle multiple rare class problems.

Ertugrul, O.F et al. (2017) have developed a method called dependent nearest neighbor (d-NN). In this method the similarity between two points i.e. the query and the sample points is obtained not only using the distance measure but also by finding the angle between them. According to this paper if two or more same distance nearest neighbor is obtained then similarity can be measured by angular information between the sample and the query; in this way same distances can be differentiated. This method of finding similarity by dependency and distance has increased the classification accuracy. However, in this case the k value has to be manually inserted.

3. Description of Broad Area

3.1 Introduction to Data mining

Immense quantity of data is produced from almost every aspect of life today which depends on information technology. Businesses such as online shopping, food chain, banks, universities, and hospitals produce data that requires analysis. These data could be in any form and are stored in large repositories. A question that arises is what can be done with this stored data? In 1980 a new revolution came up, "discovering new information from these data", which can be very useful for businesses to make smart decisions. This process of finding useful insights from a huge amount of data is called "Data Mining".

These data repositories are rich in information and when certain data mining techniques are applied on it certain hidden beneficial patterns can be obtained that can be used for business decision making. Some of the techniques for mining are clustering, classification and association. These techniques together or independently can be applied on the data repositories such as data warehouses, databases etc to obtain valuable insights.

3.2 Data mining techniques

3.2.1 Frequent pattern mining and Association

Frequent patterns (FP) are those patterns which appear more frequently than any other kind of data in a dataset. When you closely examine a transactional database of a store, you observe certain items were purchased more frequently; those patterns are called frequent patterns. In data mining frequent pattern mining is one of the more important concepts as this helps in finding association and correlation among the data. An example of frequent patterns mining is Market Basket Analysis in which customer buying habits are analysed by discovering the association between the products they buy. This discovery will help the retailer to develop many marketing strategies. For example if some items are frequently purchased together then for instance, how likely is it that a customer will buy milk if he buys bread in one trip to the store. These kinds of predictions can be made through the method of Frequent Patterns (FP) mining. The Frequent patterns mining is also helpful for data classification and clustering in data mining (Han et al. 2012) (Kumar and Upadhyay 2015).

The frequent patterns are discovered by the association rule: Following is an example for the association rule:

A customer who buys a clock also tends to buy batteries at the same time is represented by the association rule as follows:

Clock=>batteries [support=2%, confidence=60%] (Han et al. 2012)

Support and confidence are two matrices of association rule to find interestingness. A support of 2% shows that 2% of all the customers of the transactions taken under consideration for analysing will purchase clock and batteries together. Confidence 60% percent tells that 60% percentage of the customers will buy a clock and its batteries together (Han et al. 2012).

3.2.2 Clustering

In this method input-output patterns are not pre defined; hence it's an unsupervised technique. Cluster is a collection of objects which are similar to one another within a cluster and dissimilar to objects in another cluster. The process of grouping similar objects into one group is called clustering (Han et al. 2012) (Kumar and Upadhyay 2015) (Witten and Frank 2016).

There are many types of clustering method some of them are:

1. **Partition method:** In this method, the data set is divided into k parts, where k is the input parameter which is the number of partition has to be made. Then there is an iterative relocation of the objects that are made among clusters to improve the partitioning, so that each cluster will have more similar objects within the cluster than the objects in another cluster. The basic partitioning method in clustering includes k mean, k medoid (Larose, 2005) (Kumar and Upadhyay 2015).

2. **Hierarchical method:** In this cluster analysis, the objects in the dataset are organized in a hierarchy of clusters. There are two basic methods to form hierarchical clustering (Larose, 2005).

    a) **Agglomerative:** In this method, all objects in the datasets are in a separate clusters and pair of clusters merge together to form one cluster at the end. This is also called bottom up approach

b) **Divisive:** In this method, all objects are in a single cluster and the cluster is split recursively until all objects are in separate clusters. This is also known as top down approach.

3. **Density based method:** for discovering clusters of arbitrary shapes, density based method is useful. It works on the concept of density reachability and connectivity.

   Density Reachability – Two points say x and y are density reachable only if they are connected within a distance of "ε"and "y" has satisfactory amount of points in "ε" distance as its neighbors. Density connectivity means that if two points say "x" and "y" are connected then there must be a point "f" that has adequate amount of neighbors and the point "x" and "y" are "ε" distance apart from each other. If "f" has certain neighbors which are neighbors of point "x", which implies that "y" is neighbor of "x" (Han et al. 2012) (Kumar and Upadhyay 2015).

4. **Grid based method:** In this type of clustering method, the data space is divided into finite number of cells which is called grid. Clusters are discovered from this grid structure where the dense data regions are found (Han et al. 2012)

3.2.3 Classification

Classification is a data mining technique in which a classification model also known as classifier predicts the class labels. The classification process has two steps.

In the first step, a classification algorithm finds correlations between the values of target and the values of the predictors in a given dataset which is called a training data. This step is called the training or the learning step. The correlations are summarized to build a classifier model, which can then be applied to other datasets in which the class labels are unknown.

The second step is the testing phase in which a classifier model is tested by comparing the actual target values and the classifier predicted values in a set of test data (Larose, 2005).

Classification has several applications, such as scam recognition, target selling, performance prediction, manufacturing and medical diagnosis. For building a classifier model there are many methods in classification. Some of the most popular classification methods are Naive

Bayes, k nearest neighbor etc. out of which the method k-NN is considered in this study (Larose, 2005) (Witten and Frank 2016) (Kumar and Upadhyay 2015).

k-NN classifier

k-Nearest Neighbor (k-NN) classifiers are based on learning by finding similarities between the instances. The similarities are measured by a distance formula. The least distanced instances are said to be similar or close to each other. There are many distance measures such as Euclidean, Manhattan etc that are used in calculating the similarity between two instances. When an unknown query instance i.e. whose class label is not defined, when given to k-NN classifier, it calculates distances from the unknown query instances to all the other data points. The value for k is provided and the classifier selects k similar or closest instances to the unknown instance. Then the majority classes in the k similar instances are observed and whichever is the majority class label is therefore assigned to the unknown query instance. For e.g. if majority class is "male class" then the query instance is assigned as "male" (Larose, 2005).

Advantages

1. k-NN algorithm is simple and easy to implement.
2. k-NN can analyse datasets in little time which are usually handled by complex algorithms.

4 Open Issues in k-NN classifier

Even though k-NN classifier is simple to implement and a widely used classifier, it suffers a lot of drawbacks. For developing an efficient classifier, we need to solve these drawbacks. The important issues in k-NN classifier are as follows:

1. Different types of data handling issues.

   The classifiers at times are unable to handle different kinds of data such as categorical data, mixed data, noisy data, missing data which results in misclassification.
2. Finding an optimal value for k.

The selection of best k value for finding most favourable number of NN for predicting of labels more correctly is always uncertain for the user. Different k values give different predictions. So which value for k is the best value is still a problem.

3. Rare class dataset handling issue.

   Rare class or imbalanced dataset are different from standard datasets. In standard datasets, the numbers of class labels are equally distributed, but in imbalanced datasets, one of the classes will be very less in number as compared to the other class labels. So, classifying these datasets more accurately is still a challenge.

4.1 Problem Identification

4.1.1 Finding an Optimal value for k

The k is the number of nearest neighbors to an unknown instance with which we decide the class that unknown instance would fall into. It has been observed that on changing the value of k the accuracy increases and decreases accordingly. The value for k is given manually. Therefore, there is a need to obtain k value automatically (Pedrajas et al. 2015).

4.1.2 Soft boundary handling issues

The value of the k determines the number of nearest neighbors to the unknown instance. The unknown instance can be situated in any region such as middle of the class, near the boundaries of two or more classes or near the outliers or noisy data (Pedrajas et al. 2015). So there is a need to find different appropriate k values for each region for getting more accuracy in classification.

4.1.3 Finding an optimal solution for handling imbalanced dataset

Datasets such as medical diagnosis datasets, fraud detection datasets etc may have very less number of positive class instances than negative class, which is called an imbalanced datasets (Amin et al 2016).

So the model made using conventional algorithms to work on these datasets may produce biased or inaccurate predictions. The conventional k-NN model does

not consider the class distribution or proportion. They are developed to decrease error rates for improving accuracy.

4.1.4 Lazy Learning property

k-NN is a lazy learner classifier because it won't get trained from the features of the training data during the training phase instead it memorizes the training data. This can lead to misclassifications since it is not making any assumption on features and it just depends upon the majority classes of the nearest neighbors (Larose, 2005).

4.1.5 Finding appropriate distance measure issues

The k-NN classifier uses a distance measure to find the nearest neighbors and there are many distance measures such as Manhattan distance, Euclidean distance, Minkowski distance, Hamming distance and many more. The performance of the classification changes according to the distance measure used. Therefore, there is a need to know which distance measure is appropriate (Kumar and Upadhyay 2015).

4.1.6 Memory, speed and computational complexity

Each time a query instance is given to k-NN, it calculates the distance from all the data points. This property of k-NN is quite complex and a time taking task because it has to calculate the distances each time and has to store all the calculated distance values in a memory location. Therefore, the memory consumption is more.

4.1.7 Various other data handling issues

It has been observed that k-NN is less competent in handling missing data, categorical data, and noisy data.

4.2 Gaps in Literature

1. A few researchers have worked in the direction of finding optimal value for k in k-NN classification as per the existing literature.

2. Regarding the complexity issue of finding distances each time when the query instance is given, a large amount of memory is used, limited research work has been done in that direction. Solving this issue can help in increasing the speed of computation.

3. A k-NN classifier that handles imbalanced datasets and various other data such as missing values, noisy data and categorical data and multi class is yet to be perfected.

5 Objective of the study

The general objective of this research work is to discover and observe the working and performance of various classification algorithms and to improve their performance. The main aim of this novel classification approach is to classify objects with minimum misclassification rates. So, the objectives are:

1. To perform thorough investigation on various classification approaches.
2. To develop an efficient classification technique based on existing k-NN classification techniques that can handle imbalanced data and standard data.
3. To maximize the classification accuracy and minimize misclassifications.
4. To minimize computational complexity and memory usage.
5. To execute the proposed approach on data mining tools.
6. To compare the new approach with the existing approaches using data mining tools.
7. To evaluate its performance by means of classification performance matrix.
8. To enhance classification performance.

6 Methodology to be adopted

1. Literature study of working principles of various Classifiers in data mining.
2. Study and analysis on k-NN and learning about its disadvantages and challenges.
3. Understanding the various performance measures in accurate classification.

4.Comparative analysis of various existing classification approaches.

5.Detailed analysis of pre processing methods.

6.Designing new and improved pre processing and classification approach based on the existing process.

7.Comparative study on proposed and existing approaches.

8.Understanding usage of data mining tools.

7 Expected outcome of the research work

The new proposed technique based on k-NN classification technique would generate more accurate classification results. This technique will handle binary class, imbalanced class as well as multiclass problems easily with reduced computational cost. Moreover, the proposed technique would handle missing values, noisy data and categorical data much more efficiently. It can automatically find the optimal k value for minimum distance neighbourhood.

8 References

Amin, A., Anwar,S., Adnan, A., Nawaz,M., Howard, N., Qadir, J., Hawalah, A. and Hussain, A. (2016), *Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study*, Journal of IEEE access.

Ertugrul,O.F. and Tagluk,M.E. (2016), *A Novel Version of k Nearest Neighbor: Dependent Nearest Neighbor*, Applied Soft Computing Journal.

Gou, J., Du, L., Zhang, Y. and Xiong, T. (2012), *A New Distance-weighted k-nearest Neighbor Classifier*, Journal of Information & Computational Science.

Han, J., Kamber, M. and Pei, J. (2012), *Data Mining: Concepts and Techniques*, Waltham, MA, USA: Morgan Kaufmann.

Kumar, R. and Upadhyay,A.K.(2015), *Basic Concept of Data Mining*, CreateSpace Independent Publishing Platform , USA.

Larose, D. (2005), *Discovering knowledge in Data-An introduction to data mining*, A John Wiley & Sons, Inc., Publication.

Liu, H. and Zhang, S. (2012), *Noisy data elimination using mutual k-nearest neighbor for classification mining*, The Journal of Systems and Software 85 1067– 1074.

Miloud-Aouidate A. and Baba-Ali A.R. (2012), *A Hybrid KNN-Ant Colony Optimization Algorithm for Prototype Selection*, Neural Information Processing. Lecture Notes in Computer Science, vol 7665. Springer, Berlin, Heidelberg.

Nair, P., Khatri, N. & Kashyap, I. (2018), *A novel technique: ensemble hybrid 1NN model using stacking approach*. International Journal of Information Technology, pp 1–7.

Pan, Z., Wang, Y. and Ku, W. (2016), A *new k-harmonic nearest neighbor classifier based on the multi-local means*, Expert Systems with Applications, September.

Pedrajas, N.G., Castillo,J.R. and García,G.C. (2015), *A Proposal for Local k Values for k-Nearest Neighbor Rule*, IEEE Transactions on Neural Networks And Learning Systems.

Rodriguez, J., Kuncheva, L. and Alonso,C. (2006), *Rotation Forest: A new classifier ensemble method*, IEEE Transactions on Pattern Analysis and Machine Intelligence.

Witten, I. and Frank, E. (2016), *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann.

Yu, Z., Chen, H., Liu, J., You, J., Leung, H. and Han, G. (2015), *Hybrid k-Nearest Neighbor Classifier*, IEEE transactions on cybernetics 1.

Zhang, X., Li,Y., Kotagiri, R., Wu, L., Tari, Z. and Cheriet, M. (2016), *KRNN: k Rare-class Nearest Neighbor Classification*, Pattern Recognition, journal home page: www.elsevier.com/locate/pr.